

UNCLASSIFIED

AD 402 624

*Reproduced
by the*

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA



UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

6333

TM-1042/103/00

CATALOG OF ASTIA

402624

TECHNICAL MEMORANDUM

(TM Series)

ASTIA AVAILABILITY NOTICE

Qualified requesters may obtain
copies of this report from ASTIA.

This document was produced by SDC in performance of U. S. Government Contracts

Operations Research in Production
of System Training Exercises

by

K. R. Wood and H. J. Zagorsky

March 15, 1963

SYSTEM

DEVELOPMENT

CORPORATION

2500 COLORADO AVE.

SANTA MONICA

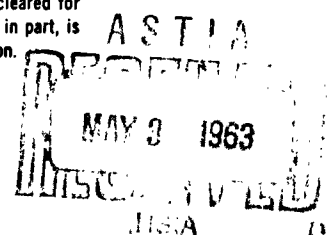
CALIFORNIA

The views, conclusions or recommendations expressed in this document do not necessarily reflect the official views or policies of agencies of the United States Government.

Permission to quote from this document or to reproduce it, wholly or in part, should be obtained in advance from the System Development Corporation.



Although this document contains no classified information it has not been cleared for open publication by the Department of Defense. Open publication, wholly or in part, is prohibited without the prior approval of the System Development Corporation.



March 15, 1963

¹
(page 2 blank)

TM-1042/103/00

TABLE OF CONTENTS

	Page
1. <u>INTRODUCTION</u>	3
2. <u>GENERAL CONCEPTION OF OPERATIONS RESEARCH</u>	4
3. <u>OBJECTIVES OF REPORTED RESEARCH</u>	6
4. <u>AUXILIARY OBJECTIVES</u>	6
5. <u>LOGICAL-NUMERICAL DESCRIPTION, FORMULATION, DEFINITION</u>	8
6. <u>SELECTION OF MODEL</u>	10
7. <u>SIMPLIFICATION OF MODEL</u>	11
8. <u>APPLICATION OF MODEL--REGRESSION MODELLING</u>	12
9. <u>RESULTS OF THE OPERATIONS RESEARCH</u>	17
10. <u>CURRENT STATUS OF ESTIMATING EQUATION DEVELOPMENT</u>	21
APPENDIX I. FORECASTING AND ESTIMATING EQUATIONS DERIVED	23
APPENDIX II. MODEL FOR PRODUCTION FLOW	27
APPENDIX III. OPTIMIZING SELECTION OF VARIABLES FOR ESTIMATION AND FORECASTING	29
BIBLIOGRAPHY	35

OPERATIONS RESEARCH IN PRODUCTION OF SYSTEM TRAINING EXERCISES

by

K. R. Wood and H. J. Zagorski

1. INTRODUCTION

The following is a report of research conducted during the past three years in production scheduling of air defense system training exercises. The system training exercise or problem* is a simulation of aircraft movement, of aircraft offensive and contrasurveillance activity, and of certain defense-system reactions. System training problem production is therefore concerned with production of Air Defense System simulation inputs and accessory materials for a program of system training.

Regarded separately, training, simulation, and production are areas of extensive enquiry and development. They must, furthermore, be considered jointly in the development, installation, maintenance and operation of a production system for system training, for dynamic simulation. Moreover, logical integration must include real-time integration. General definition of the training problem must respond to changing training requirements in different environments and to evolving conceptions of these requirements. Operational definition of the training problem must consider current and prospective capabilities in simulation and production. Finally, complete specification

*Training exercise" and "training problem" are used interchangeably in referring to the product.

March 15, 1963

4

TM-1042/103/00

of the individual training problem and its simulation requisites is an iterative process, proceeding in part from materials already in process of production for that or for related training problems.

In view of the foregoing, in view of the product's evolving and subtle nature as a simulation for training, of its iterative generation, and of its multi-discipline development and production, opportunities for operations research are to be anticipated and are pursued. Aspects of this unique production technology are so diverse, certain questions and results of their study so specific, and other considerations so general and challenging, that a variety of reasonable definitions of operations research may find exemplification.

2. GENERAL CONCEPTION OF OPERATIONS RESEARCH

In general, the ultimate object of operations research is improvement of operations. Objective measures of such improvement are desirable, but may be only partially attainable. Even where achieved, objective measures will rarely supply all the information required for decision. Consensus of opinion, expert and considered opinion, and other information will also supply bases for decision. The extent, however, to which quantitative and objective analysis and modelling of operations can suggest sounder courses of action will certainly be of interest to operations research and administrative personnel alike.*

*Churchman, C. W., Ackoff, R. L., and Arnoff, E. L. Introduction to Operations Research. New York: Wiley and Sons, 1957, Parts I and X.

In rough chronological outline, research of operations implies (1) an operation and objectives for its research; (2) logical-numerical information describing, defining and formulating the operation in its context; (3) one or more abstractions or conceptions (models) of the operation; (4) logical, symbolic, intuitive, heuristic, or mathematical operation upon or application of the model; (5) derivation of hypotheses--suggestions, insights, or implications--from consideration of the model; (6) testing of the hypotheses arising from modelling; and (7) communication, narration, and description of findings as a basis for action. Additionally, an implementation or action stemming from the research might require (8) assisting in planning or guiding the action, (9) observation of results and perhaps iteration or recycling of (1) to (9), inclusive, until a satisfactory convergence ("closure") obtains or proves unattainable.*

As a science, operations research acknowledges (1) a world of observed phenomena, on the one hand, and (2) abstractions or conceptions ("models"), on the other.** The need for dynamic, continuous, reciprocal mapping or interchange between phenomenon and abstraction, and the need for alternate use of induction-to-model and deduction-to-phenomenon, as seen in the previous paragraph, would seem inevitable. In a given experiment or area of research, a critical factor for valid abstraction may lie unnoticed in either phenomenon or model, in microcosm or macrocosm. The effective joining of observations

* Churchman, Ackoff, and Arnoff, op. cit., Parts II, III and IX.

** Ibid., Chapter 1.

and of model constitutes the science. While history has recorded eras of pure reason and eras of empiricism, operations research must respect both for the achievement inherent in their synergism.

3. OBJECTIVES OF REPORTED RESEARCH

The original objective of the reported research was to assist in the production scheduling operation by making possible a more realistic anticipation of the flow of training problems through the production system. This objective was temporarily hardened to that of forecasting more accurately than previously possible, completion dates of particular production activities on each training problem.

A second objective, deriving naturally from interest and activity centered on the first, was to examine and improve estimation of computer-time requirements for individual training problems.

A third objective (actually rounding out the first) was to simulate the flow of training problems through the production system, projecting the simulation into the future in order that states of overload and underload might be anticipated.

A fourth objective was to ascertain what feasible and current changes in schedule might improve future overload and underload situations in the various stages of production.

4. AUXILIARY OBJECTIVES

The above objectives were the primary objectives of research by the authors. They are for the most part statements, by personnel responsible for production,

of what is needed. As is nearly always the case in operations research, auxiliary and allied objectives arose for consideration. Through auxiliary objectives, original objectives are analyzed, separated into logical components, rigorously specified, and perhaps abstracted, in order that useful principles, concepts, laws, rules, or expedients may be indicated and applied.

A first auxiliary objective was concerned with the area of work measurement, with seeking definition of a unit of effort to be applied in assessing the magnitude of production operations. The realities of production scheduling had long before compelled an expedient by production personnel for estimating production units (or production unit weights). The production unit is a useful index of the effort entailed in the computer phase of production for a given training problem. The work measurement objective would imply improvement of the production-unit-weight estimator, if possible, and its generalization to other phases of the production process.

A second auxiliary objective became one of adapting and improving applied regression analysis (or more properly least -squares analysis, or data processing*) to permit the indispensable iterative interaction of empirical statistical modelling and of research insight, and to permit empirical confirmation of regression-analysis results and their utility.

*Tukey, J. W. The Future of Data Analysis. The Annals of Mathematical Statistics, 33 (1), March 1962, pp. 1-67.

A third auxiliary objective was the motivation and implementation of data collection, the supplying of impetus for definition, recording, preservation, retrieving, rearranging, and summarizing of data pertinent for operations research.

A fourth auxiliary objective arose with the shift in emphasis from Gantt or bar-chart production scheduling** to network methods of production scheduling and control. This objective was to relate or contrast some of the prerequisites for and benefits from network applications on the one hand, and bar-chart applications on the other.

5. LOGICAL-NUMERICAL DESCRIPTION, FORMULATION, DEFINITION

When the currently reported research began, a weekly schedule was in use, showing the anticipated completion dates of production stages for each training problem throughout the production year. Nearly one hundred training problems were produced annually. Production time varied with the nature of the product, from a few weeks to several months. The production process, as represented on the weekly schedule, was comprised of a dozen or more stages, some, however being of only a few days duration.

Since the product varied widely in a dozen or more important characteristics, and in fact required some element of iterative definition in the very process of production, prediction of production time was inherently difficult. Rules of

** Reinfeld, N. V. Production Control. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1959, pp. 139-140

thumb were of necessity prevalent, there being considerable difference in the rules and the variables they emphasized. It was generally agreed, however, that improvement was required in the anticipation of completion dates.

At the outset of research, the twelve or more production stages shown on the production schedule were consolidated into five consecutive stages, each requiring roughly one-fifth of the total production time and having a somewhat natural demarcation from the other stages. Production Department personnel later evolved, in more detail, a network representation of the production process, and employed PLAN, a PERT-like production scheduling and control method.

Three classes of variables were distinguished: product variables, system-facility variables, and system-state variables. It was assumed that these variables influenced the production time required in each stage (in varying emphasis from stage to stage) and that their values would permit anticipation of this time.

An inevitable obstacle at the very outset of research was the lack of numerical information, of data-taking and retrieving capability. In this new production technology, initial energies had of necessity been directed to production methods. Increase in production load then brought a need for scheduling--for data and information upon which to base the scheduling operation. It became apparent that a subsystem for data definition, recording, storing, retrieving, arranging, analyzing, and presentation, was indispensable--and almost without precedent for the time and technology.

Without benefit of previous data and analysis, the very definition of data, the definition and selection of variables, was a purely expert-judgment or

educated-guess operation. Knowledge of and insight into the production system, discussion with system personnel, pilot analysis of fragmentary and even questionable data, all contributed to definition and selection of data for collection. Design of a data system must acknowledge, as does design of experiment, the source and definition of the data, and its ultimate use. The collection of useful data required, among other things, convincing others of the need for data, training personnel, coordinating and mutual consulting with system personnel, analyzing the production system logically for convenient, dependable, and meaningful sources of data, the devising of data forms and the handling and analysis of data.

The process of achieving a data system stressed communication in a large, involved, and necessarily changing system. Different, conflicting definitions were detected in this evolving technology and resolved by the team approach. Conflicting records and record-keeping methods were encountered and examined, record gaps revealed, and uniform methods and formats adopted. The purpose of the research was constantly reviewed and communicated in order that definition and collection of data be appropriate.

6. SELECTION OF MODEL

The many and varied characteristics of the product, the previous use in the production system, of intuitive, weighted-variable estimation, the emphasis upon estimation and prediction for planning purposes, the prevalent discussion of the relative importance of variables, the lack of a priori functions or laws, and the existence of computer programs for statistical analysis, made reasonable

March 15, 1963

11

TM-1042/103/00

the choice of multiple-regression and factor-analysis methods in attempting to meet the stated objectives.

For each of k successive production stages, a single-regression function was to be obtained for predicting processing time required in that stage. A rough model, or the components of a rough model of production flow, would then consist simply of a single-regression function for each of the k successive stages of production. For the j th stage, the j th regression function would be employed to estimate the processing time required for the j th stage from training-problem characteristics and from system descriptors. A computer program would be written to produce such estimates, to accept tentative system entry dates, and to project for several months in advance, the weekly production status anticipated for each training problem. The program would also compile for each week the resultant work load in each stage of production.

7. SIMPLIFICATION OF MODEL

System-facility and system-state variables were ultimately dropped from the model partly for reasons of constraints on the data available from the minimal data system then possible, partly because facilities such as number of computers or number of computer shifts merely offset, to a degree, the system state, and partly because linear multiple-regression analysis, consistent with the foregoing, failed to reveal definite (additive) contribution to the success of prediction by such currently available variables. The use of a queue variable to predict completion date in a regression function seemed nevertheless a concept worth noting for possible future consideration.

March 15, 1963

12

TM-1042/103/00

8. APPLICATION OF MODEL--REGRESSION MODELLING

The central problem in application of regression modelling was determination of the mathematical form of the regression function. When, in such application, the functional form is predesignated and assumptions of the general linear hypothesis are reasonably tenable, classical point-estimate method can indeed be useful. (Here the roles of statistics as separate discipline and as "servant to the sciences" have much in common.) As may frequently be the case, however, the form of the ultimate regression function was by no means apparent a priori. Questions were involved of mathematical type of function, selection of variables, selection of their transformation, number of terms, and (for polynomials) degree. Thus, the assumption of a unique, predesignated functional form in fixed-value variables having neither error nor sampling distribution, giving rise to normally, independently and homogeneously distributed errors, was untenable.* Empirical evidence against adoption of this compound assumption is afforded in many research studies in which the dependent variable is approximated equally well, essentially, by more than one, by linearly independent combinations of the so-called independent variables. A uniquely best or least-squares solution is typically indeterminate. Still other questions concerning heterogeneity of data and curvature of regression make even more precarious the routine adoption of the classical multiple least-square procedure.

Customarily, when parameters are too many for the number of independent equations available, constraints are invoked to obtain a unique solution. It

*Wood, K. R. Multiple Regression Analysis, Synergism of Practice and Theory. SDC document N-18085, April 24, 1962. This document is an internal, unpublished communication and is not appropriate for release outside the corporation.

is, however, the source of these constraints that may be given more consideration--an interdisciplinary consideration. If they are to be invoked more or less arbitrarily, from the mathematical point of view, the constraints may as well have a meaningful, subjective purpose--and be supplied by the researcher. The mathematics has nothing to lose, so to speak, while the research area has the distinct possibility, even the necessity, of gain. This principle is fully illustrated in the application of factor analysis or characteristic vector analysis in which a vector subspace is established by least-squares, but the choice of basis within that space is not arbitrary, being in part, at least, subjective and meaningful in terms of information beyond immediate data and information. There seems to be no reason whatsoever for not applying the same concept, of a multiple-solution space, to the area of multiple-regression analysis. Application of this concept, in fact, makes possible an interface between mathematical aspects and research area considerations in modelling. Information in the data and information transcending the data are thereby permitted an operationally defined and imperative joint consideration.

In the currently reported study in applications of conventional multiple-linear-regression analysis, the existence of a multiple-solution space of least-squares-equivalent solutions was again and again demonstrated. The least-squares solution was very rarely, if ever, unique, and constraints in the form of extra criteria were therefore indispensable. These could be supplied only from the area under study.

In employing data analysis, the research investigator is typically interested in criteria of utility, feasibility, parsimony, and interpretability.

Since these criteria are essentially irrefutable in import and yet often difficult to define and incorporate in a composite and objective mathematical criterion for selection of regression function, they must be applied as subjective and meaningful considerations within the multiple-solution space--in the choosing of a unique or near-unique regression function, for example. Ideally, then, a multiple-solution space would first be established primarily by some mathematical criterion (least squares for example); a coordinate system, a pattern or basis of solution points or vectors spanning this solution-parameter space would be produced; and the research investigator, thus assisted, would select one or more solutions which, in some degree of relative emphasis, are useful, feasible, parsimonious, and meaningful.

To illustrate, in the typical application of multiple linear regression, in which more than one independent linear combination of the independent variables afford essentially the same success in approximation, orthogonality of solutions (of coefficients employed in linear approximations) may, for convenience, be stipulated. The orthogonal solutions may be produced in the order of their merit by least-squares criterion. The better, essentially equivalent solutions, as basis vectors, can then be combined (components of them can be taken) by the investigator or the research team in search of one or more solutions that appear particularly interpretable, useful, feasible, and parsimonious. Thus, from a mathematically delineated space of solutions, expert knowledge of the research area would be employed to select one or more solutions of particular importance with respect to considerations currently incapable of mathematical formulation.

As previously acknowledged, the principle of a multiple-solution space has long been observed operationally in applications of factor analysis or characteristic vector analysis. Inherent in the original intent and very formulation of factor analysis is the concept of a vector subspace, least-squares determined, in which the basis, however, may be selected in various ways. The literature on multiple-regression analysis, on the other hand, implies for the most part a mathematically unique solution, a point estimate. When practical experience, proceeding on the assumption of uniqueness, yields absurd or uninterpretable answers stemming from near vanishing of determinants, from ill conditioned matrices, from near dependence of "independent" variables, from heterogeneity of data, and from curvilinearity, confusion is inevitable until the concept of the multiple-solution space is again acknowledged.

The first and most natural expedient commonly adopted, and one utilized in the current study as well, is the use of several reasonable or expert-judgment combinations of variables in the multiple least-squares approximation. Results are examined not alone for success of approximation, but for reasonability--interpretability, feasibility, utility, and parsimony. One or more solutions are accordingly selected.

A second expedient, having an extensive literature, is that of variable selection (or in the tests and measurements area, "test selection"). Mathematically stated, the object of variable selection is as follows: From r vectors, obtain that linear combination of only k vectors by least-squares criterion in terms of which still another vector is best approximated. In so far as the authors are aware, this mathematical problem is as yet unsolved

March 15, 1963

16

TM-1042/103/00

except, of course, by exhaustive (and often quite unfeasible) inventory of all (\bar{k}) solutions and inspection for the best (or better).

Acknowledging both the mathematical and operational implications of the foregoing problem, two computer programs, A34 and A28, were developed for the purpose of finding superior least-squares-equivalent solutions, A34 for linear approximation, A28 for joint second-degree polynomial approximation (including the linear as a special case). Both programs are part of the SDC statistical library, and have been employed on a variety of projects.

Program A34 produces superior linear-approximation solutions by a gradient method described in FN-6622/000/00, "Multiple Regression With Subsetting of Variables," dated 11 June 1962. Albeit heuristic, the method has produced a number of instances in which a confirmed best pair of vectors did not include the best single vector for use in linear approximation. Unlike many variable-selection procedures, the method of A34 does not necessarily employ all of a selected subset of $k-1$ vectors in compiling a selected subset of k vectors for use in linear approximation. In some instances, however, conventional selection methods produced a set of variables superior to the set selected by A34. This occurs when a member of the best k variables is itself very nearly linearly dependent upon remaining variables. It is believed that the method can be modified and improved. In its present form it compares favorably with and augments other methods.

Program A28 permits fitting a general quadratic form by least squares, but provides for its expression in orthogonal-polynomial components, each component being simply a product of linear forms. The method of A28 is sometimes described

as "product-of-linear forms" regression. Near linearity of data results in a near-constant value for one of the two linear forms, and the approach therefore to conventional linear regression. The first derived product-of-linear-forms polynomial is "best" in the least-squares sense, the second derived being second best, and so on--yielding a canonical form for the general quadratic form fitted by least squares. An unsolved problem is the assignment of degrees of freedom to the successively derived and component products of linear forms.

9. RESULTS OF THE OPERATIONS RESEARCH

With respect to the first objective, specific, empirical equations were developed, tested, and applied in predicting elapsed production time in each stage. Changes, trends and perturbations in the system inevitably had their effect, but a useful stability was nevertheless achieved. From the many variables involved, a half dozen were found especially useful and meaningful. The choice between (1) a single, but relatively involved, formula for all types of training problems, and (2) a simpler but different formula for each of several logical classes of products was resolved in favor of the latter. Even the crude queue variables available gave evidence that somewhat better definition, collection, and application of variables might well permit improvement by the use of system-state (load) variables. Eventual incorporation of such variables in a regression model was considered realistic.

With respect to the second objective, an intensive comparison of the relative effectiveness of a regression function with a production unit gave clear evidence in favor of the regression function--even when the production

unit was itself given the benefit of least squares fitting. An interesting sidelight was the frequent failure of intuitive estimates to make use of a constant term--in spite of the prevalence of such concepts as initial costs or overhead. The regression function derived contained a positive and reasonable constant term.

With respect to the third objective, a program, B319 was written and applied in prediction of stage workload for several months in advance. The validity of the projection with respect to major peaks and valleys was not only confirmed, but comparison with the forecasts being obtained with the bar-chart method were quite favorable. Furthermore, the graphic representation of prospective workload, obtainable from B319, made assessment of the system state much more convenient and realistic. The effect of production start dates upon imbalance was much better realized.

With respect to the fourth objective, a load-balancing program or supplement to B319 was not achieved. Perhaps a factor was the general acceptance of a stipulated or predesignated start date, with little allowance for variation in either direction. Questions of definition of imbalance and of method of balancing were, however, very considerable. A rough outline of a heuristic load-balancing algorithm was nevertheless developed, which borrowed from the very useful concept of the "response surface" and the gradient. An observation was to the effect that operational mathematics often exacts of itself an optimum solution (in a very literal and rigorous sense), while the operation (system) may aspire only to a better and perhaps more immanent solution.

With respect to the first auxiliary objective, a fundamental and parsimonious reconstruction of the production unit was the primary accomplishment. As an ingeniously and intuitively constructed work measure, augmented from time to time, the production unit had served a very useful purpose. Underlying the production unit were found two basic variables, length of the training problem-input (PI) tape and length of the training film. Upon elaborations of these two basic variables there had been appended some separately reasonable, "allow-so-much-for" rules which, however, gave no confirmable increase in predicting capability. An interaction of empirical analysis and a priori considerations, mutually self stressing and sustaining, demonstrated that simpler and more meaningful approximation was achievable with PI tape length and film length. Other data available permitted the estimation of planning time required in problem production. Whereas machine-time variables or analogues were the meaningful variables in the computer area of production, number of flights (of different types) gave rise to useful and meaningful estimation in the planning area.

With respect to the second auxiliary objective, statistical support programs, A34, A28, and A52 were developed, programmed, and employed. The concept of the multiple-solution space was evolved. The parallel between the Kelley-Salisbury concepts from the tests and measurements field and the concept of relaxation from the numerical analysis field was pursued in program A28. Since parsimony is perhaps easier to define than utility, feasibility, and interpretability, but may, indeed, be in some degree concomitant with the latter, the relaxation method is employed in A28 to search

the multiple-solution space for parsimonious solutions. Conventional joint polynomial second-degree approximation is nonetheless possible with the program.

With respect to the third auxiliary objective, concerted effort toward prediction, estimation, and production-flow simulation stimulated cooperative efforts and interaction in the area of data definition, collection, and retrieval. The data problem is, of course, receiving more and more attention in science and industry, but deserves still more attention and effective treatment. Very good convergence to sound definition of useful data was obtained in interactions of operations research and operational personnel, and very significant strides were taken in obtaining meaningful data. The minimal data system was successful particularly in demonstrating the potential realization from operational definition of information.

With respect to the fourth auxiliary objective, experience and reflection served to answer, in part, questions frequently raised with respect to network scheduling and bar-chart scheduling. Clearly (by definition) bar-chart scheduling assumes a one-path network for each of several products on which activity proceeds more or less in parallel. Many applications of network scheduling, on the other hand, presuppose only one very large and complex product. The contrast is that of several or many successive-stage tasks represented in bar-chart scheduling with a single, very complex task represented by the network. The necessary and sufficient operations of the complex task can, of course, be portrayed more faithfully in the network and the planning made more specific and effective. If the large task having the network of activity is performed only once or not often, data may be unavailable and estimation of

processing times for the various component operations may be one of pure judgment, good or bad. To the extent that several network tasks may be involved, that the networks are similar, that each network is approximated by a succession (single path) of subnetworks, and that data become available, successive-stage modelling may be very useful as a condensation of the more detailed model. Particularly for the purposes of appraising general facility requirements and the impact of peaks and valleys, the more condensed and simpler modelling may prove useful, the search for valid parsimony having a history of payoff. The history of factor analysis affords an example in which, typically, many seemingly independent variables or considerations are actually empirically dependent, and prediction, therefore, need be concerned only with a few basic considerations.

10. CURRENT STATUS OF ESTIMATING EQUATION DEVELOPMENT

Until the current fiscal year (FY 1963), the data used to ascertain functional relationships between job variables and production costs had been gathered by catch-as-catch-can methods. No means existed whereby either the cost or the work characteristics of STP (System Training Program) problems could be appropriately assembled for systematic analysis. Of necessity, variables chosen for analysis were those which were readily and economically available to the research investigators.

With the installation of a cost-recording system (April 1962) and the establishment of a project work group to integrate and expedite the construction of a system data base, (August 1962) the complexion of the estimating

March 15, 1963

22

TM-1042/103/00

equation development has changed considerably. For example, a labor cost analysis is currently being conducted in nine operational areas in system training problem processing. This analysis is expected to yield preliminary insights regarding the differential cost factors in these areas. When the full system for assembling and processing the data base swings into action, it is expected that most of the significant cost areas in problem processing will be described equationally and factorially.

A year ago, the refinements in estimating toward which OR is now working were impossible. Though all the credit for this apparent change is not due solely to operations research (a number of sound management decisions have helped considerably), the sustained application of operations research principles and practices through thick and through thin is believed to have had a strong influence on the growth and advancement being shown.

APPENDIX I

FORECASTING AND ESTIMATING EQUATIONS DERIVED

A. Computer Time

Some a priori considerations and empirical analysis indicate different classes of products and a different estimating function for each class. In Class I are training problems made for smaller manual air defense units in areas such as Germany, Spain, Alaska and Hawaii. In Class II are training problems made for larger manual air defense units, primarily in Canada. In Class III are training problems made for larger SAGE air defense units, primarily in the United States. Finally, in Class IV are training problems made for smaller SAGE air defense units in the United States.

For the Class I training problems, the following equation was selected from a number of alternatives:

$$H = .34F + 5.6$$

where H is the number of computer hours estimated for the problem and F is the total number of training film hours to be delivered to the exercising unit. This equation assumes approximately 130 flights per problem unit. Since there is some contribution to the computer time from this characteristic, estimates provided by the above equations are adjusted by .004 hours for each flight deviating from the average of 130.

For Class II problems, the following equation was selected from among the alternatives:

$$H = .69F + 3.9$$

where H and F are defined as before. This equation was based on an average of 370 flights per problem unit. The adjustment for deviation from this average is .023 computer hours per flight.

For Class III problems, the following equation was selected:

$$H = 10.6T - 56,$$

H being defined as before, and T being the total number of training-magnetic-tape hours to be delivered to the exercising unit. This equation is based on an average of 530 flights per problem unit. The adjustment for deviation from this average is .010 computer hours per flight.

For Class IV problems, the following equation was selected:

$$H = 2.7T + 10.4,$$

H and T being defined as before. This equation was based on an average of 270 flights per problem unit. The adjustment for deviation from this average is .025 computer hours per flight.

The above equations were delivered to the production-control process for use in forecasting and scheduling computer time. Their joint use with traditional estimating methods has served to stimulate interest in data definition and sources, in recognition of the need for good estimating factors or variables, in the saving of estimates and errors of estimate as a basis for improving estimating methods, and in the comparison and combination of various estimating methods.

B. Man Time Estimation

The data available were in the area of training-problem planning. Training-problem planners are individuals who set up the specifications for

training materials and control these specifications throughout the production process. They also perform associated creative work dealing with training aids and descriptive materials. For this activity the following equations was derived:

$$M = 1.56S + 330,$$

where M is the estimated number of man hours required to plan each problem unit and S is the number of flights which must be controlled within that existing unit. When there are several units in a problem, M must be calculated separately for each unit and summed over all units to get an aggregate estimate of planning time.

C. Elapsed Time Forecasting

Equations were derived to forecast the time required for a product (training problem) to reach a given phase of production. Five phases of production were defined:

- Phase I Computer Planning and computer inputs manuscripting.
- Phase II Computer input preparation and initial processing.
- Phase III Computer output review and modification of computer inputs.
- Phase IV Computer problem production proper.
- Phase V Problem finishing and shipping.

Again there existed a need for classification of product. One class consists of training problems where several different problem units are produced separately and assembled as a large-scale problem. The equations derived for this class of product were:

Phase I Completion Date = Start Date + 63 days + 4P

Phase II Completion Date = Start Date + 114 days + 10P

Phase III Completion Date = Start Date + 131 days + 10P

Phase IV Completion Date = Start Date + 169 days + 17P

Phase V Completion Date = Start Date + 198 days + 17P.

P represents the number of problem (areas) units entailed.

A second major class of product consists of smaller training problems built essentially for one manual-air-defense-problem unit. These equations were:

Phase I Completion Date = Start Date + 20 days + F

Phase II Completion Date = Start Date + 47 days + 2F

Phase III Completion Date = Start Date + 61 days + 2F

Phase IV Completion Date = Start Date + 76 days + 2F

Phase V Completion Date = Start Date + 98 days + 2F

F represents the total number of film hours in the problem unit.

APPENDIX II

MODEL FOR PRODUCTION FLOW

A. Definition of Model

For a given product or item (training problem) having a given start date for production, the equations in Appendix I, Section C permitted estimation of completion date for each phase of production. The traditional production unit, P, served as a rough measure of the magnitude of the production effort entailed for the product. A computer program was written to employ equations of the type given in Appendix I, Section C, to obtain estimates of completion dates, and to sum the values of P in each phase of production for each future week of production. This sum of P for each phase of production was plotted graphically for future weeks, to anticipate load peaks and valleys.

B. Algorithm for Improving Load Balance (Proposed)

A load-balancing algorithm was proposed, which would shift start dates conservatively, in an effort to improve the work-load balance obtained by the above mentioned computer program. An average or normal workload would be agreed upon or computed for each phase of production. A weight, W, would be adopted for each phase of production, for the purpose of conveying the relative import of imbalance among the phases of production. A measure of imbalance, the W-weighted squared-deviation-from-normal load would be summed over weeks and over phases. This statistic, Z, would be computed for any proposed or modified schedule. The computer would be employed to slip (delay start of production for) each product, alone and in turn, one week. For each

March 15, 1963

28

TM-1042/103/00

of the resulting schedules (like the original except for slipping one item), Z would be computed. That particular single change of schedule accomplishing the greatest decrease in Z would be tentatively adopted. This process would be repeated until only trivial decrease in Z are realized. The final schedule obtained would then be rescaled in its real-time dimension, to bring the total elapsed time within a reasonable period.

APPENDIX III

OPTIMIZING SELECTION OF VARIABLES FOR ESTIMATION AND FORECASTING

A. Definition of Problem

1. For Linear Model

Which k of n available reference vectors define a k -space in which another given vector has maximum length (projection)? Or, with which k of n available column vectors can the corresponding values of another column vector best be approximated, by least-square criterion? Is there a best set of k ? Are there nearly-best sets? How does one choose from the best and nearly best sets? Must all combinations, n things taken k at a time, be examined? And, for the realities of application, what integral value shall k have?

2. For Polynomial Models

What combinations of variables and of terms yield best or nearly best least-squares approximations? How does one choose from the best and nearly best combinations?

B. Operational Treatment of Problem

1. For Linear Model

If for any of the k variables of a subset, SS , employed in a linear approximation, there can be substituted variables of the full set with consequent improvement of approximation, SS is, by definition, not a best subset. A necessary (but not sufficient) condition, therefore, for a best subset is that any partition of the subset into two complementary partial subsets will find the partial subsets best mutual complements.

The method described below employs an approximate method in finding a usually superior complementary partial subset for a given partial subset. For an original and perhaps arbitrary partial subset (PSS1), a complementary partial subset (PSS2) is found. In turn, for PSS2 a new PSS1 is found, and so on--experience with an actual computer program (SDC Library Program A34) indicating that mutual complementation ordinarily results, the partial subsets eventually renominating each other. It can be shown that this is not sufficient for a best subset. Furthermore the method of optimal complementing is only approximate. Nevertheless, in very extensive, practical experience the method has compared quite favorably with other methods, frequently finding confirmed superior pairs of variables which do not include the best one.

It is assumed that a partial subset of variables has been employed in a least-squares approximation. For each of the independent variables available, an element is computed for a vector, \underline{G} , derived below. Those variables having elements of highest absolute value in the vector \underline{G} , are selected for the new, complementary partial subset. (From the nature of the least-squares process, elements of \underline{G} , for the variables currently employed in the approximating partial subset, are of zero value.) Depending upon the scaling of the independent variables, the elements of \underline{G} may be regarded as elements of a gradient or as elements of a constrained exact total differential, for the coefficient of multiple determination with respect to change in variable regression coefficients. Except for the empirical evidence gained in actual computer application, the appeal of the method resides largely in these two properties of \underline{G} .

Let the matrix-vector product, Xb , represent a linear combination of the column variables of X (all available independent variables). Let the column variable Y be approximated by Xb (with arbitrary elements of b), and let $E = Y - Xb$ represent the column of errors resulting from the approximation of Y by Xb . The elements of the column vector, b , are taken as variable, any least-squares derivations, at present, being special cases.

Regard the X and Y column variables as having a mean of zero and a vector magnitude of unity. The scalar product, E -transpose E , or $E'E$, represents the error variance. This variance is equal to $1 - 2r'b + b'Rb$, where r is $X'Y$ (the column vector of validities or correlations of Y with the X 's) and R is $X'X$, the matrix of intercorrelations for the independent variables.

Decrease and minimization of $E'E$ is equivalent to increase and maximization of $D = 2r'b - b'Rb$, a general expression for the coefficient of determination (least squares derived or not). When the elements of the vector b are least squares derived (except perhaps for certain elements being specifically fixed at zero in value), $b'Rb$ and $r'b$ alike are the coefficients of determination. The above expression for D then becomes $D = 2D - D$. The original problem may be stated, however, as one of maximizing $D = 2r'b - b'Rb$ with respect to the elements of vector b , subject to the condition that a prescribed number of elements of the vector b are zero in value.

$$(1) \Delta D = 2r'(\Delta b) - 2b'R(\Delta b) - (\Delta b)'R(\Delta b).$$

Let $g' = r' - b'R$, gradient of D with respect to b scaled by $1/2$.

$$(2) \Delta D = 2g'(\Delta b) - (\Delta b)'R(\Delta b).$$

For maximum ΔD , the partial derivatives of ΔD , with respect to elements of Δb , must be zero:

$$(3) 2g' - 2(\Delta b)'R = 0. \quad g'(\Delta b) = (\Delta b)'R(\Delta b).$$

$$\Delta D = g'(\Delta b) = (\Delta b)'R(\Delta b).$$

If, for $\Delta b = b_2 - b_1$, b_1 is current least-squares solution, with zero-valued elements for variables not involved in the approximation, then $b_1'g_1 = 0$, and

$$(4) \Delta D = g_1'(b_2 - b_1) = b_2'g_1.$$

If the b_2 elements are least-squares coefficients, B_j , for all variables, and ΔD is therefore the difference between current D and maximum D ,

$$(5) \Delta D = B'g_1, \text{ the scalar product of } B \text{ and of } g.$$

If each original variable, x_j , is regarded as being scaled by B_j in its contribution, $B_j g_j$ to $B'g$, the elements $B_j g_j$ are those of the gradient, \underline{G} of D , with respect to the unit-valued coefficients of the variables, $B_j x_j$.

2. Polynomial-Regression Models--An Algorithm for Least-Squares or Near-Least-Squares Determination of a Product of Linear Forms

Supposing the matrix X to have a column pseudovector (having always the value of unity), let the matrix-vector product, Xc , represent a

linear combination of the column variables of X (all variables available for use in an approximating formula). Likewise, let the matrix-vector product, Xd , represent a second linear combination of the column variables of X . Let x_c represent the i th value of the column variable Xc , and x_d represent the i th value of the column variable Xd . Let the product of x_c and x_d represent an approximation of the i th value y of a column variable, Y . Let $e = y - (x_c)(x_d)$ represent the resulting error of approximation. Furthermore let E represent the column variable of e -values, and suppose that the product, E -transpose by E , or $E'E$, the sum of squared errors, is to be reduced if not actually minimized.

In the attempt to preserve simplicity, suppose that the variable Y is at first approximated by its mean value, all values of the vectors, c and d , being zero except for the c -coefficient and d -coefficient of the pseudovalue, their product being the mean value of Y . The change in $E'E$ is readily derivable for any change in c or d . In the interest of promoting simplicity of approximation, that single change of one c -coefficient or of one d -coefficient may be made, which maximally reduces $E'E$. Having made such a change, of course, one variable has been introduced into the approximating function, $(x_c)(x_d)$. Again that single change of one c -coefficient or of one d -coefficient may be made, which maximally reduces $E'E$. This process is readily continued (on the modern computer) until $E'E$ is no longer appreciably reduced. Typical experience with this algorithm indicates that values of the c - and d -elements attain stable values as $E'E$ is no longer appreciably reduced. But further and trivial

March 15, 1963

34

TM-1042/103/00

reductions in $E'E$, in the effort to achieve a least-squares solution, may be accompanied by radical changes in certain elements of c and d , indicating of course that the least-squares solution has many competitors, differing only trivially in $E'E$ but markedly in values of c and d . Typically, however, $E'E$ will have achieved a nearly minimal value when the elements of c and of d have stabilized at long-enduring values, a number or many of the initial zero values being preserved.

Bibliography

1. Arnquist, C. W., Dembiczak, C. M., Wood, K. R., and Zagorski, H. J. Forecasting Computer Time from STP Problem Characteristics. SDC document SD-3384.* August 4, 1960.
2. Lintner, R. M. Use of MAPS in the Data Processing Department. SDC document N-18631.* July 23, 1962.
3. McCornack, R. L., Villone, L., and Wood, K. R. Multiple Regression With Subsetting of Variables. SDC document FN-6622/000/00. June 11, 1962.
4. Weinwurm, G. F. ADD Contributions to Symposium on Operations Research at SDC. SDC document N-18731/002/00.* August 9, 1962.
5. Wood, K. R. A Conjecture on the Composition of the Best Subset of p Variates Available for Use in Linear Approximation (Multiple Regression). SDC document SD-3322.* January 4, 1960.
6. Wood, K. R. Hypothesis Seeking--A Proposal for Facilitating Polynomial Multiple Regression When the Number of Variables is Large. SDC document SD-3389.* August 26, 1960.
7. Wood, K. R. A System of Statistical Programs for Linear and Curvilinear Regression and Factor Analysis. SDC document D-3478.* January 9, 1962.
8. Wood, K. R. and Zagorski, H. J. Fact Sheet on MAPS. SDC document SP-352 May 24, 1961.
9. Wood, K. R. and Zagorski, H. J. How to Schedule by Computer...Use MAPS. Aerospace Management, 5(2), February 1962, pp. 36-37.
10. Wood, K. R. and Zagorski, H. J. MAPS, A Mathematical Management Technique. SDC document SP-634/000/01, January 1962.

* This document is an internal, unpublished communication and is not appropriate for release outside the Corporation.

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
OPERATIONS RESEARCH IN PRODUCTION
OF SYSTEM TRAINING EXERCISES.
TM-1042/103/00, by K. R. Wood,
H. J. Zagorsky. 15 March 1963,
35p., 10 refs.

Unclassified report

DESCRIPTORS: Air Defense Command.
Operations Research.

Reports on the research conducted
during the past three years in

UNCLASSIFIED

production scheduling of air
defense system training exercises.
Describes the system training
exercise (or problem) as a
simulation of aircraft movement, offensive,
and contrasurveillance activity, and
of certain defense-system reactions.
States that system training problem
production is therefore concerned with
production of Air Defense System
simulation inputs and accessory
materials for system training.

UNCLASSIFIED

UNCLASSIFIED